MULTISTRATEGY ENSEMBLE REGRESSION FOR MAPPING OF BUILT-UP DENSITY AND HEIGHT WITH SENTINEL-2 DATA

Christian Geiß, Henrik Schrade, Patrick Aravena Pelizari, and Hannes Taubenböck

German Remote Sensing Data Center (DFD), German Aerospace Center (DLR), 82234 Weßling-Oberpfaffenhofen, Germany; christian.geiss@dlr.de, henrik.schrade@dlr.de, patrick.aravenapelizari@dlr.de, hannes.taubenboeck@dlr.de

ABSTRACT - In this paper, we establish a workflow for estimation of built-up density and height based on multispectral Sentinel-2 data. To do so, we render the estimation of built-up density and height as a supervised learning problem. Given the rational level of measurement of those two target variables, the regression estimation problem is regarded as finding the mapping between an incoming vector, i.e., ubiquitously available features computed from Sentinel-2 data, and an observable output (i.e., training set), which is derived over spatially limited areas in an automated manner. As such, training sets are automatically generated from a joint exploitation of TanDEM-X mission elevation data and Sentinel-2 imagery, and, as an alternative, from cadastral sources. The training sets are used to regress the target variables for spatial processing units which correspond to urban neighborhood scales. From a methodological point of view, we introduce a novel ensemble regression approach, i.e., multistrategy ensemble regression (MSER), based on advanced machine learning-based regression algorithms including Random Forest Regression, Support Vector Regression, Gaussian Process Regression, and Neural Network Regression. To establish a robust ensemble, those algorithms are learned with a modified version of the AdaBoost.RT algorithm. However, to reliably ensure diversity between single boosted regressors, we include a random feature subspace method in the procedure. In contrast to existing approaches, we selectively prune non-favorable regressors trained during the boosting procedure and calculate the final prediction by a weighted mean function on the residual models to ensure enhanced accuracy properties of predictions. Finally, outputs are concatenated into a single prediction with a decision fusion strategy. Experimental results are obtained from four test areas which cover the settlement areas of the four largest German cites, i.e., Berlin, Hamburg, Munich, and Cologne. The results unambiguously underline the beneficial properties of the MSER approach, since all best predictions were obtained with a boosted regressor in conjunction with a decision fusion strategy in a comparative setup. The mean absolute errors of corresponding models vary between 3–16% and 1–5.4m with respect to built-up density and height, respectively, depending on the validation strategy, size of the spatial processing units, and test area. Also in a domain adaptation setup (i.e., when learning a model over a source domain and applying it over a geographically different target domain) numerous predictions show comparable accuracy levels as predictions obtained within a source domain. This further underlines the viability to transfer a model and, thus, enable a substitution of the training data in the target domains.

Index Terms – Sentinel-2, TanDEM-X, urban morphology, built-up height estimation, built-up density estimation, regression models

1. Introduction

Global urbanization processes and population growth reshape the landscapes of our planet. Built environments develop from sparsely populated settlement areas to urban agglomerations with millions of people (Taubenböck *et al.*, 2012; UN, 2017). In this context, the systematic and continuous characterization of built environments is an essential step for enabling various analyses and applications. These comprise tailored analyses related to urban planning and environmental management (e.g., Heiden *et al.*, 2012; Huang *et al.*, 2017), as well as dedicated applications associated to e.g., natural hazard risk (e.g., Geiß and Taubenböck, 2013; Pittore *et al.*, 2017), or energy-related assessments (e.g., Ratti *et al.*, 2005; Geiß *et al.*, 2011), among others. Thereby, constituent properties of urban morphology such as *built-up density* and *height* can serve as proxies and descriptive features to support aforementioned analysis and applications. As such, *built-up density* represents one of the most relevant descriptive, explanatory, as well as normative measure in urban research (e.g., Taubenböck *et al.*, 2016), whereas the vertical dimension of built environments as approximated by *built-up height* must be considered to enable a holistic assessment of environmental relationships found in urban areas (e.g., Berger *et al.*, 2013).

To quantify and map *built-up density* and *height* properties, Earth Observation (EO) data were identified as a valuable source of information, since EO systems represent data collection mechanisms for continuous measurements in time and space. Past studies frequently relied on digital surface models (DSM) (from e.g., LiDAR measurements) and optical imagery (from e.g., WorldView, GeoEye etc.) with a very high spatial resolution (VHR) to resolve and analyze the objects of built environments such as buildings. For instance, Yu *et al.* (2010) derive a normalized DSM from LiDAR measurements to compute density-related measures such as building coverage ratio and floor area ratio, among others. Likewise, Gonzáles-Aguilera *et al.* (2013) deploy DSM information from VHR LiDAR data for derivation of geometric properties (height, area, and volume) as well as density attributes (building coverage ratio and floor area ratio) of buildings, land lots, and urban units. Very recently, attempts were followed to alleviate these data requirements and estimate the DSM from VHR multispectral imagery using advanced supervised learning techniques. Thereby, a learning machine is trained on scenes where both the DSM and optical data are available to establish an image-to-DSM translation rule (Mou and Zhu, 2018; Ghamisi and Yokoya, 2018). Supervised learning techniques were also deployed for estimation of built-up density from VHR optical imagery. In this manner, Zhang *et al.* (2018) learn models from labeled samples on multiple image features to map built-up density. To alleviate the restrictions that are associated with the proper collection of prior knowledge (i.e., compilation of a training set), Heinzel and Kemper (2015) establish an unsupervised workflow based on VHR multispectral imagery for a joint description of built-up areas according to maximum building size, heterogeneity of the building size, and built-up density. To this purpose, they use operations from mathematical morphology (Soille, 2004) on the VHR multispectral imagery. In order to map built-up areas with a high accuracy, Liu *et al.* (2019) exploit multi-view data from Ziyuan 3 multispectral satellite imagery and tailored angular difference features (Huang *et al.*, 2018) within an unsupervised mapping scheme.

However, the usage of VHR data can still hamper deployment capabilities for very large areas due to general data availability, monetary costs, and demanding processing requirements. When aiming at spatially continuous analysis and assessment approaches for very large areas such as countries, continents, or even the globe, these kinds of data can still represent a major limitation nowadays. Nonetheless, recent EO systems feature a remarkable tradeoff between a high spatial resolution and large-area coverage. Regarding elevation information, the TanDEM-X mission (TDM), which is a spaceborne radar interferometer, provides a global DSM with an unprecedented pixel spacing of 0.4 arcseconds (~12 meters) (Krieger *et al.*, 2007; Zink *et al.*, 2014). Regarding optical imagery, ESA's Sentinel-2 satellites provide superspectral imagery with a spatial resolution of 10 meters for the bands covering visible light and near infrared and feature a field of view of 290 kilometers (Drusch *et al.*, 2012).

Sentinel-2 imagery was already deployed to estimate the degree of imperviousness (Lefebvre *et al.*, 2016; Xu *et al.*, 2018). However, a joint derivation of *built-up density* and *height* was just recently proposed by Geiß *et al.* (2017, 2019) based on an integrative analysis of TDM and Sentinel-2 data. As such, these data sets allow for a unique mapping of urban morphology around the globe for large areas. Notably, Sentinel-2 data are provided free of charge to the public via a data hub, which is accessible online (Copernicus, 2018). In addition, the imagery can also be queried and processed via the Google Earth Engine (Gorelick *et al.*, 2017). In contrast, although the data exists consistently for the whole globe (Wessel *et al.*, 2018), the accessibility of TDM data is currently limited to 100 000 square kilometers per data proposal for scientific applications. To alleviate this limited accessibility, here, we aim to develop an approach to substitute the required TDM elevation data for mapping of *built-up density* and *height*. To allow for truly large-area application, we combine TDM elevation data and Sentinel-2 imagery and render the mapping of *built-up density* and *height* as a supervised learning problem. Given the rational level of measurement of these two variables to be predicted, we generate a regression model based on areas where data from both TDM and Sentinel-2 are available. With it, we aim to estimate the target variables for areas where only Sentinel-2 data is available. Consequently, the regression estimation problem is regarded as finding the mapping between an incoming vector $\mathbf{x} \in \mathbb{R}^F$, i.e., ubiquitously available features F computed from Sentinel-2 imagery, and an observable output $y \in \mathbb{R}$ (i.e., *built-up density* and *height*) from a given pool of labeled training samples $\boldsymbol{X} = \left\{\left(\mathbf{x}_i, y_i^{density}, y_i^{height}\right)\right\}_{i=1}^{l}$, i.e., automatically derived according to the workflow proposed in Geiß *et al.* (2017, 2019) over spatially limited areas where data from both TDM and Sentinel-2 are available. In addition, we evaluate the incorporation of geospatial vector data, i.e., level of detail 1 (LoD-1) building geometries from cadastral sources, which also enable the computation of *built-up density* and *height* and, thus, allow compiling a training set.

From a methodological point of view, we introduce a novel ensemble regression approach based on advanced machine learning-based regression algorithms. In the past, algorithmic regression models (Breiman, 2001) were successfully deployed in the context of remote sensing to achieve a high level of predictive accuracy (e.g., Esch *et al.*, 2009; Leinenkugel *et al.*, 2011; Verrelst *et al.* 2012; Aghighi *et al.*, 2018). To further enhance accuracy and robustness of predictions, ensemble learning strategies were also implemented (Okujeni *et al.*,

2017; Feilhauer et al., 2015). The goal of ensemble learning methods is to combine different, probably suboptimal models into a single model with enhanced predictive accuracy properties. To do so, such methods establish a set of models, which were learned individually for a given problem. Subsequently, this set of models (i.e., the ensemble) is combined to establish the final prediction (Mendes-Moreira *et al.*, 2012). Thereby, different strategies can be followed to impose model independency. The training samples which are presented to the learning algorithm can be modified by drawing randomly training sample subsets with replacement, i.e., "bagging" (Breiman, 1996) or by reweighting iteratively the training samples, i.e., "boosting" (Freund and Shapire, 1996). In addition, models can be learned based on training sets with randomly selected features, i.e., random subspaces (Ho, 1998). Also, multiple learning algorithms can be trained. In order to address challenging classification problems, such kind of approaches (i.e., multiple classifier systems) are particularly popular (Du *et al.*, 2012). Finally, predictions of the different learning algorithms can be combined based on decision fusion strategies (Polikar, 2006). For regression problems the combination scheme must take the ratio scale of measurement into account to allow for superior solutions with respect to the individual model predictions. For instance, so-called meta-classifiers such as stacked generalization (Wolpert, 1992), which relearn a model using also the outputs of previously learned models, enabled beneficial predictive accuracies previously.

To address all of the aforementioned aspects, we establish an innovative ensemble regression approach. The term multistrategy ensemble regression (MSER) is subsequently used when referring to our technique, since we uniquely combine multiple ensemble learning methods within an integrative framework (Webb and Zheng, 2004). To account for the No-free-Lunch-Theorem, which states that there is no algorithm that induces the most accurate learner in any domain all the time (Wolpert, 1996), we jointly consider the model outputs of advanced machine learning-based regression algorithms including Random Forest Regression (RFR), Support Vector Regression (SVR), Gaussian Process Regression (GPR), and Neural Networks (NN). To establish a robust ensemble, those algorithms are learned with a modified version of the AdaBoost.RT algorithm (Solomatine and Shrestha, 2004). However, to reliably ensure diversity between single boosted regressors, we include a random feature subspace method in the procedure. In contrast to related approaches (Garcia-Pedrajas and Ortiz-Boyer, 2008; Wang and Ma, 2011), we selectively prune non-favorable regressors trained during the boosting procedure and calculate the final prediction by a weighted mean function on the residual models to ensure enhanced accuracy properties of predictions. Finally, outputs are concatenated into a single prediction by application of a decision fusion strategy. In this context, we evaluate the performance of stacked generalization and local selection. The latter partitions the feature space in decision regions according to the locally most accurate single regression algorithm (Bruzzone and Melgani, 2005). Besides, we implement an exhaustive feature calculation module. Previous studies already underlined the capability of multispectral imagery to sophistically describe buildings and built-up structures (Zhang *et al.*, 2017). Consequently, we compute an exhaustive number of spectral features, features related to mathematical morphology (Soille, 2004), as well as texture measures derived from the gray–level co–occurrence matrix (GLCM) (Haralick, 1979) from Sentinel-2 imagery. To select beneficial features for regression, we built upon previous work and draw a set of features

from the initial feature vector from a multi-view perspective (Aravena Pelizari *et al.*, 2018). Finally, domain adaptation (Tuia *et al.*, 2016) capabilities of the approach (i.e., learning a model on a source domain and reliably applying it on a geographically different target domain) are ensured in an unsupervised way by implementing a histogram matching procedure, which is eventually run before the feature calculation module to address a possible covariate shift.

Generally, the spatial resolution properties of the data hamper analyses on individual building level. The pixel spacing of 0.4 arcseconds for TDM data and ten meters for multispectral Sentinel-2 imagery can exceed the extent of the objects of interest (i.e., buildings). As a consequence, we work on an aggregated spatial level, i.e., we establish spatial processing units in terms of rectangular grid cells to compute *built-up density* and *height* thereof. The spatial processing units correspond to urban neighborhood scales. In this manner, experimental results are provided from computations regarding the settlement areas of the four largest German cities: Berlin, Hamburg, Munich, and Cologne. Thereby, we limit the data processing to settlement areas according to the so-called Global Urban Footprint (GUF) layer. This is a binary mapping scheme which discriminates "built-up" and "non built-up" areas globally with a high spatial resolution (Esch *et al.*, 2012, 2017).

The remainder of the paper is organized as follows. Section 2 gives an overview of the proposed methodology and section 3 is used to present the deployed data sets and explain the experimental setup. Section 4 provides experimental results and validation efforts. Concluding remarks and an outlook to future work are given in section 5.

2. Proposed Methodology

An overview on the input data and processing steps for mapping of *built-up density* and *height* based on Sentinel-2 imagery is provided in **Fig. 1**. First, spatial processing units are created by aggregating the GUF data set to rectangular grid cells. The grid cells are derived from Sentinel-2 imagery. The spatial processing units constrain analyses explicitly to settlement areas. Subsequently, a training set is compiled by either relying on an automatic workflow which estimates *built-up density* and *height* based on areas where both TDM elevation data and Sentinel-2 imagery are available (Geiß *et al.*, 2017a, 2019). In addition, we evaluate the incorporation of geospatial vector data from cadastral sources for this purpose (details can be found in the experimental setup section).

A feature calculation module is implemented which contains the computation of spectral features, features based on mathematical morphology, and texture features. It is intended to provide an exhaustive description of built environments for model learning and prediction (section 2.1). To incorporate solely beneficial features in the ensemble regression, we establish a multi-view filter-based feature selection approach (section 2.2). The actual MSER approach foresees the learning of regression models based on multiple advanced machine learning-based algorithms (section 2.3), a boosting procedure with random feature subspace (section 2.4), and concatenation of multiple models into a single prediction with suitable decision fusion strategies (section 2.5). Eventually, domain adaptation capabilities of the

approaches are addressed by a histogram matching procedure, which is directly run before the feature calculation module (section 2.6).

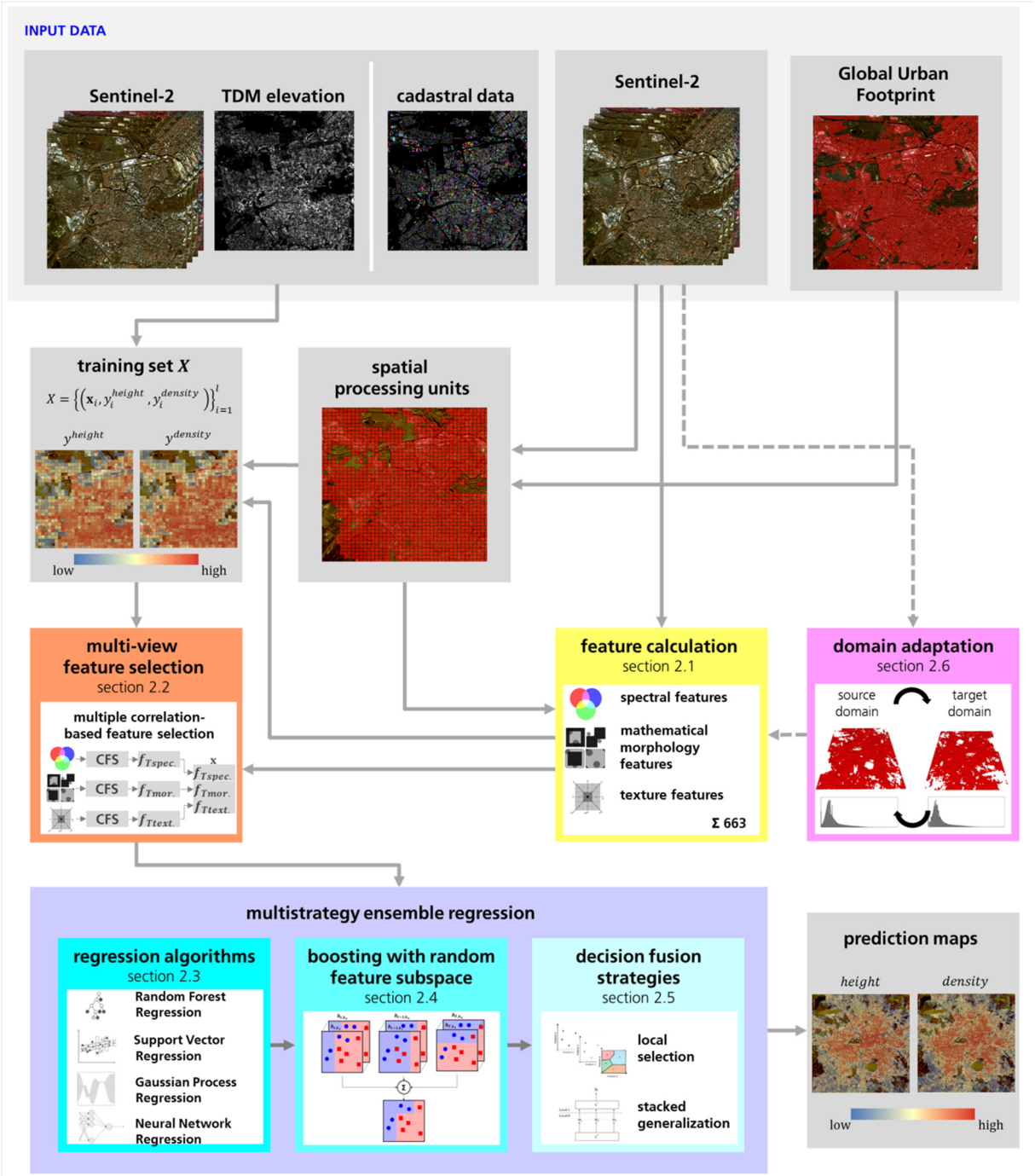

Fig. 1. Overview on the input data and processing steps for mapping of *built-up density* and *height*. Details on the methodological aspects are given throughout the subsections of section 2.

2.1 Feature calculation

Features from three different categories are considered to describe the urban morphology in this study. We deploy Sentinel-2's blue (490 nm), green (560 nm), red (665 nm), and NIR (842 nm) bands with a spatial resolution of 10m (Drusch *et al.*, 2012) individually for feature calculation to maintain and exploit the maximum amount of available information.

1) *Spectral features*

Buildings and built-up structures show frequently a distinct pattern in multispectral imagery. They have a high absolute reflectance in the visible spectrum and appear brighter than their surroundings due to the frequent strong contrast between bright roofs and dark adjacent shadows (Pesaresi and Gerhardinger, 2011). We exploit the non-transformed spectral information by aggregating the image bands to the spatial processing units according to different measures of central tendency and spread (i.e., mean, median, standard deviation, variance, minimum, maximum, and interquartile range (IQR)) (**Table I**). Moreover, spectral differences can provide additional information (Geiß *et al.*, 2015). As a well-known example, the normalized difference vegetation index (NDVI) (Rouse *et al.*, 1973) provides discriminative information regarding vegetated and non-vegetated areas. Consequently, we compute a set of six spectral ratios (Zhang *et al.*, 2017) to emphasize spectral dissimilarities, as well as a brightness measure, which represents the mean pixel values regarding the bands covering the visible and NIR spectrum. These measures and all subsequent measures are aggregated to the spatial processing units with the mean function.

TABLE I. FEATURES COMPUTED FROM SENTINEL-2 IMAGERY

Feature category	Feature name and description	No.
spectral	$blue\ band_{mean,median,std.dev.,var.,min.,max.,IQR}$ $green\ band_{mean,median,std.dev.,var.,min.,max.,IQR}$ $red\ band_{mean,median,std.dev.,var.,min.,max.,IQR}$ $NIR\ band_{mean,median,std.dev.,var.,min.,max.,IQR}$	28
	$NDVI_{mean}$ $normalized\ NIR\ green\ index_{mean}$ $normalized\ NIR\ blue\ index_{mean}$ $normalized\ red\ green\ index_{mean}$ $normalized\ red\ blue\ index_{mean}$ $normalized\ green\ blue\ index_{mean}$ $brightness_{mean}$	7
mathematical morphology	$DMP\ based\ on\ Opening\ and\ Closing_{mean}^{R,G,B,NIR}$ $DMP\ based\ on\ Opening\ and\ Closing\ by\ Reconstruction_{mean}^{R,G,B,NIR}$	208
	$MBI_{mean}\ using\ DMP\ based\ on\ Opening^{brightness}$ $MSI_{mean}\ using\ DMP\ based\ on\ Closing^{brightness}$ $MBI_{mean}\ using\ DMP\ based\ on\ Opening\ by\ Reconstruction^{brightness}$ $MSI_{mean}\ using\ DMP\ based\ on\ Closing\ by\ Reconstruction^{brightness}$	4
texture	$GLCM\ mean_{mean}^{R,G,B,NIR}$ $GLCM\ variance_{mean}^{R,G,B,NIR}$ $GLCM\ homogeneighty_{mean}^{R,G,B,NIR}$ $GLCM\ contrast_{mean}^{R,G,B,NIR}$ $GLCM\ dissimilarity_{mean}^{R,G,B,NIR}$ $GLCM\ entropy_{mean}^{R,G,B,NIR}$ $GLCM\ angular\ 2nd\ moment_{mean}^{R,G,B,NIR}$ $GLCM\ correlation_{mean}^{R,G,B,NIR}$	416
		Σ 663

2) *Features based on mathematical morphology*

Image descriptors based on mathematical morphology (Soille, 2004) are particularly suitable to both highlight local extrema of the spectral signal and extract shape-related properties of

the objects of interest in an exhaustive manner (Geiß *et al.*, 2016). We built upon basic erosion and dilation operations to compute differential morphological profiles (DMPs) (Pesaresi and Benediktsson, 2001) using concatenated opening and closing operations with an increasing size of the structuring element (SE) (more details on the parameterization can be found in the experimental setup section). In addition, DMPs of opening and closing by reconstruction operations are also considered to synergistically preserve the initial shape properties of the image objects. To establish an exhaustive description, DMPs are computed individually for all four images bands. Besides, we computed variations of the Morphological Building Index (MBI) and the corresponding counterpart, i.e., the Morphological Shadow Index (MSI) (Huang and Zhang, 2012; Huang *et al.*, 2014). These indices exploit the fact that the relatively high reflectance of roofs and adjacent shadow areas induce a high local contrast of buildings and built-up structures. First, a dedicated brightness image is computed by recording the maximum value with respect to the visible bands. The NIR band is neglected here since the visible bands contribute most severely to the spectral properties of buildings and built-up structures (Pesaresi *et al.*, 2008). This brightness image is used to compute DMPs of both opening and closing operations and opening and closing by reconstruction operations. Here, the opening operations are deployed for the MBI (i.e., emphasizing the bright image structures) and the closing operations are used for the MSI (i.e., emphasizing the dark image structures), respectively.

3) Texture features

Quantification of surface texture can yield powerful discriminative properties regarding built environments (e.g., Zhang *et al.* 2018). Consequently, a set of texture features based on the GLCM (Haralick, 1979) is considered. These features were shown to carry supplementary information when the spectral resolution is limited and the ground sampling distance is much smaller than the objects of interest. Such a situation can be frequently found in VHR multispectral imagery (Geiß *et al.*, 2017b). However, this situation also occurs for our setting when characterizing built-up structures based on Sentinel-2 data. Consequently, we selected eight measures from the set of 14 originally proposed GLCM measures, since some are strongly correlated with each other and we aim for a possibly minimal computational burden. The co-occurrence frequencies of grey-levels are typically quantified in symmetric matrices for pixels in spatial proximity along 0°, 45°, 90°, or 135°, respectively. Rotation-invariance of a GLCM feature can be obtained by summing up the directional GLCMs before computation of the feature (Stumpf und Kerle, 2011). We deploy two first order statistics and six second order statistics (**Table I**), which showed favorable performance properties in studies analyzing imagery of built environments (Pacifici *et al.*, 2009; Huang *et al.*, 2014) (more details on the parameterization can be found in the experimental setup section).

Given the different feature categories and various window sizes of the deployed spatial features, overall, each spatial processing unit is represented by a 663-dimensioanl feature vector before feature selection.

2.2 Multi-view feature selection

High dimensional data tend to exhibit a large amount of redundancy as well as irrelevant noise. Thereby, an increased dimensionality of the feature space promotes the susceptibility of learned inference models to suffer from the Hughes phenomenon (Hughes, 1964) and overfitting. Feature selection (FS) addresses these problems. FS generally leads to more compact feature sets, which facilitate data interpretation and result in more cost-effective models (Guyon, 2003).

The use of different feature types that represent a different perspective (also referred to as view) on the inference target, each with a specific physical meaning and distinct statistical properties (e.g., spectral and spatial features), preserves the exhaustiveness of the entire feature space. Complementary subspaces are imposed that generally show a high heterogeneity and rather low redundancy among each other. In contrast, features within a subspace are likely homogeneous and show high redundancy (Chen *et al.*, 2017). Several studies have shown that FS approaches which account for multiple existing views perform better than single-view approaches (Zhao *et al.*, 2017; Chen *et al.*, 2017; Aravena Pelizari *et al.*, 2018).

With regard to supervised FS, methods can be categorized into *filters* (Duch, 2006), *wrappers* (Kohavi and John, 1997), and *embedded approaches* (Lal *et al.*, 2006). Unlike wrappers and embedded approaches, filters work independent of the learning algorithm. This property makes them flexible to use and therefore particularly convenient to be deployed within an ensemble learning approach (section 2.3). Besides, filter algorithms feature usually less computational complexity and demand less computation time (Kohavi and John, 1997; Lal *et al.*, 2006). Hence, we deploy a filter technique and follow the *multiple Correlation-Based Feature Selection* (MCFS) approach (Aravena Pelizari *et al.*, 2018) which is a modification of the *Correlation-Based Feature Selection* (CFS) technique (Hall, 2000) to establish multi-view capabilities. CFS originally evaluates individual subsets of the entire feature vector F based on an evaluation criterion that favors subsets with a high feature-target variable correlation and low feature-feature inter-correlation:

$$m_S = \frac{k\overline{r_{fy}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}}, \tag{1}$$

with m_S being the merit of subset S, k is the number of features in the subset, $\overline{r_{fy}}$ refers to the average feature-target variable correlation, and $\overline{r_{ff}}$ represents the average feature-feature inter-correlation.

In contrast to this, MCFS first decomposes F into predefined disjoint typological sub-feature spaces (f_T) and subsequently applies CFS on the single f_T. Each f_T refers to an individual view and is assumed to imply relatively homogeneous orders of scales in its feature manifestations and feature-target variable correlations. Accordingly, the resulting subset for classification $MCFS(F)$ based on n typological sub-feature spaces f_{T_i} $i \in [1, \dots, n]$ can be defined as:

$$MCFS(F) = \left(CFS_{f_{T_1}}, CFS_{f_{T_2}}, \dots, CFS_{f_{T_n}}\right), \tag{2}$$

with $CFS_{f_{T_i}}$ denoting the CFS subsets of f_{T_i}. To speed up the application of CFS, a best-first strategy (Russel and Norvig, 2010) is deployed for searching the feature subspaces.

We applied MCFS considering the spectral bands, spectral ratios, morphological profiles, MBI and MSI features, and texture features (**Table I**) as individual views, i.e., compiling a reduced feature vector $\hat{F}$ which internalizes five views.

2.3 Machine learning-based regression algorithms

The deployed regression algorithms are briefly described in this subsection. Details on parametrization and hyperparameter tuning can be found in the experimental setup section.

1) Random Forest Regression

RFR builds upon the bootstrap aggregation principle and randomly draws a subset of n features to create splits for each tree of each learner (Breiman, 2001; Liaw and Wiener, 2002). As a result, different models for different learners are established. Subsequently, the total trees grown to construct the prediction model $\hat{p}_Q$ are concatenated:

$$\hat{p}_Q = \frac{1}{Q}\sum_{q=1}^{Q} T(\mathbf{x}_*, \Theta_q) \tag{3}$$

where Q represents the total number of trees, $T(\mathbf{x}_*, \Theta_q)$ is the qth tree, $\mathbf{x}_*$ corresponds to the instance which requires a prediction value, and Θ_q represents a random vector established for the qth tree independent of previous ones (Aghighi *et al.*, 2018).

2) Support Vector Regression

SVR builds upon the popular SVM framework (Cortes and Vapnik, 1995). SVR defines a linear model over samples which are mapped with a nonlinear kernel function to a higher dimensional space. When using Vapnik's ε-insensitive cost function, SVR determines weights $\mathbf{w}$ by minimizing the regularized functional:

$$\min_{\mathbf{w},\xi_i,\xi_i^*,b}\left\{\frac{1}{2}\|\mathbf{w}\|^2 + C\sum_i(\xi_i + \xi_i^*)\right\} \tag{4}$$

subject to

$$y_i - (\langle\phi(\mathbf{x}_i),\mathbf{w}\rangle + b) \leq \varepsilon + \xi_i \qquad \forall i = 1,\ldots,n \tag{5}$$

$$(\langle\phi(\mathbf{x}_i),\mathbf{w}\rangle + b) - y_i \leq \varepsilon + \xi_i^* \qquad \forall i = 1,\ldots,n \tag{6}$$

$$\xi_i, \xi_i^* \geq 0 \qquad \forall i = 1,\ldots,n \tag{7}$$

where ξ_i and ξ_i^* are positive slack variables, which quantify the distances of the labeled training samples outside of the ε-insensitive tube to the border of the tube. C establishes a trade-off between the flatness of the function and the tolerance to experienced errors during training. The estimation function is given by:

$$f(\mathbf{x}_*) = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)K(\mathbf{x}_i,\mathbf{x}_*) + b. \tag{8}$$

where α_i are Lagrange multipliers, K being a kernel function, and b the bias term in the regression (Geiß *et al.*, 2016).

3) Gaussian Process Regression

Similar to SVR, GPR is a kernel-based approach (Rasmussen and Williams, 2006). To infer an unknown functional relationship from a training data set, a Gaussian Process prior is placed upon the latent function and a Gaussian prior is used for each latent noise term to constrain the possible forms of the unknown function first. Subsequently, the priors are updated under consideration of the training data to establish a posterior GPR model (Camps-Valls *et al.*, 2016):

$$f(\mathbf{x}_*) = \sum_{i=1}^{n} \alpha_i K(\mathbf{X}_i, \mathbf{X}_j) + \omega_0. \tag{9}$$

where α_i is the weight associated to each training sample, K is a function which evaluates the similarity between the training set $\mathbf{X}_i$ and test set $\mathbf{X}_j$, and ω_0 corresponds to the bias term.

4) Neural Networks

NN represent a (potentially fully) connected structure of neurons which are organized in layers. A single neuron establishes a linear regression before establishing a nonlinear activation function $f(\cdot)$. We built a multi-layer perceptron with the input layer, one hidden layer, and the output layer, respectively. Neurons of the different layers are interconnected with links, i.e., weights: a neuron j in layer $l+1$ yields

$$x_j^{l+1} = f\left(\sum\nolimits_i w_{ij}^l x_i^l + w_{bj}^l\right) \tag{10}$$

where w_{ij}^l are the weights connecting neuron i in layer l to neuron j in layer $l+1$, and w_{bj}^l corresponds to the bias term of neuron j in layer l (Verrelst *et al.*, 2012). Here, the sigmoid function was used as $f(\cdot)$.

2.4 Boosting with random feature subspace

To establish a robust ensemble, the previously described algorithms are learned with a modified version of the AdaBoost.RT algorithm (Solomatine and Shrestha, 2004). AdaBoost.RT builds upon AdaBoost (Freund and Shapire, 1996) which was developed to improve the prediction accuracy of weak classifiers with a prediction accuracy slightly better than random guessing. Building upon a given weak classifier, AdaBoost builds an ensemble by training the classifier multiple times, $t = \{1, \ldots, T\}$, whereas the training data is iteratively sampled with replacement between the steps. After each iteration, the weights for the initial sampling distribution are updated based on the performance of the preceding weak hypothesis, giving greater weight to samples that were misclassified in the previous step (Kummer and Najjaran, 2014). Finally, AdaBoost combines the results of the weak classifiers into a single prediction based on a weighted majority vote in which the weight of each weak classifier is a function of its accuracy (Freund and Schapire, 1997).

The initial AdaBoost formulation focused on binary classification. Subsequently, the approach was extended to deal with multiple classes (AdaBoost.M1) and regression problems (AdaBoost.R) (Freund and Schapire, 1996, 1997). Solomatine and Shrestha's (2004) AdaBoost.RT algorithm builds upon AdaBoost.M1 and pursues the general AdaBoost procedure: samples are drawn based on a weight vector D_t which is updated after each iteration. However, in contrast to classification settings where predictions are either true or false, predictions in regression are rarely perfectly accurate and, naturally, discrepancies between the predicted and actual values are inevitable. In order to classify a predicted value as correct or incorrect the absolute relative error (ARE) is calculated for each prediction:

$$ARE_t(i) = \left|\frac{f_t(\mathbf{x}_i) - y_i}{y_i}\right| \tag{11}$$

where $f_t(\mathbf{x}_i)$ represents the current hypothesis (i.e., regression estimate) and y_i is the correct numerical value of the corresponding training sample. Subsequently, the error of each prediction is compared with the relative error threshold ϕ, which was introduced by Solomatine and Shresta (2004) as an additional hyperparameter. By counting the correct and incorrect predictions, the error rate ε_t of $f_t(\mathbf{x})$ is calculated as follows:

$$\varepsilon_t = \Sigma D_t(i);\ i: ARE_t > \phi. \tag{12}$$

The error rate again serves as input for the computation of the weight updating parameter β:

$$\beta_t = \varepsilon_t^2. \tag{13}$$

At the end of an iteration step the weight vector D_t is updated by β:

$$D_{t+1}(i) = \frac{D_t(i)}{Z_t} * \begin{cases} \beta_t, if \left|\frac{f_t(\mathbf{x}_i) - y_i}{y_i}\right| \leq \phi \\ 1, otherwise \end{cases} \tag{14}$$

where Z_t is a normalization factor to model D_{t+1} as distribution. The final output is generated by calculating the weighted average:

$$f_{fin}(\mathbf{x}) = \frac{\sum_{t=1}^{T}\left\{\left(\log\frac{1}{\beta_t}\right) * f_t(\mathbf{x})\right\}}{\sum_{t=1}^{T}\left(\log\frac{1}{\beta_t}\right)} \tag{15}$$

AdaBoost.RT was conceptualized for improving weak regressors. However, the predictions of the deployed regression algorithms are far from random guessing after the hyperparameter tuning. Consequently, the diversity of the boosted regressors is not sufficient and the results produced by the single regressors provide insufficient variation if solely different sets of training data are provided to the tuned regressors. Thus, to reliably ensure diversity between single boosted regressors, a random feature subspace method is included into the procedure. The random subspace method (RSM) was presented by Ho (1998) and aims at increasing the diversity of classifiers by varying the composition of features during training. In related

approaches, the combination of AdaBoost and RSM were successfully tested (Garcia-Pedrajas and Ortiz-Boyer, 2008; Wang and Ma, 2011) for classification problems.

In order to adapt the AdaBoost.RT algorithm to non-weak regressors, we prune non-favorable regressors trained during the boosting procedure. As such, the procedure was extended by three additional steps: i) a feature subspace of size s is created before the sampling procedure by drawing features randomly without replacement; ii) in addition to the ARE the root–mean–square error (RMSE) of each single boosted regressor with respect to the training samples is calculated; iii) an additional relative threshold σ is introduced. Based on the calculated RMSE, only the $T * \sigma$ best boosted regressors are chosen for generating the final output according to eq. (15).

2.5 Decision fusion strategies

To concatenate various model predictions into a single final estimate, we implement two decision fusion strategies.

1) Local selection

Local selection was introduced to the remote sensing community by Buzzone and Melgani (2005). The idea is to partition the feature space into various regions of specific decisions defined by the most accurate local model prediction (Giacinto and Roli, 2001). To this purpose, in a first step, individual models are trained and the most accurate local predictions are determined with a test set (**Fig. 2a**). Subsequently, the feature space is divided in decision regions which will induce predictions for unseen samples according to the locally most accurate model determined in the previous step. The division of the feature space can be obtained via the *k*-nearest neighbor decision rule. Finally, unseen samples will be located in feature space and the actual estimation of values will be made by the regression model associated to the corresponding region.

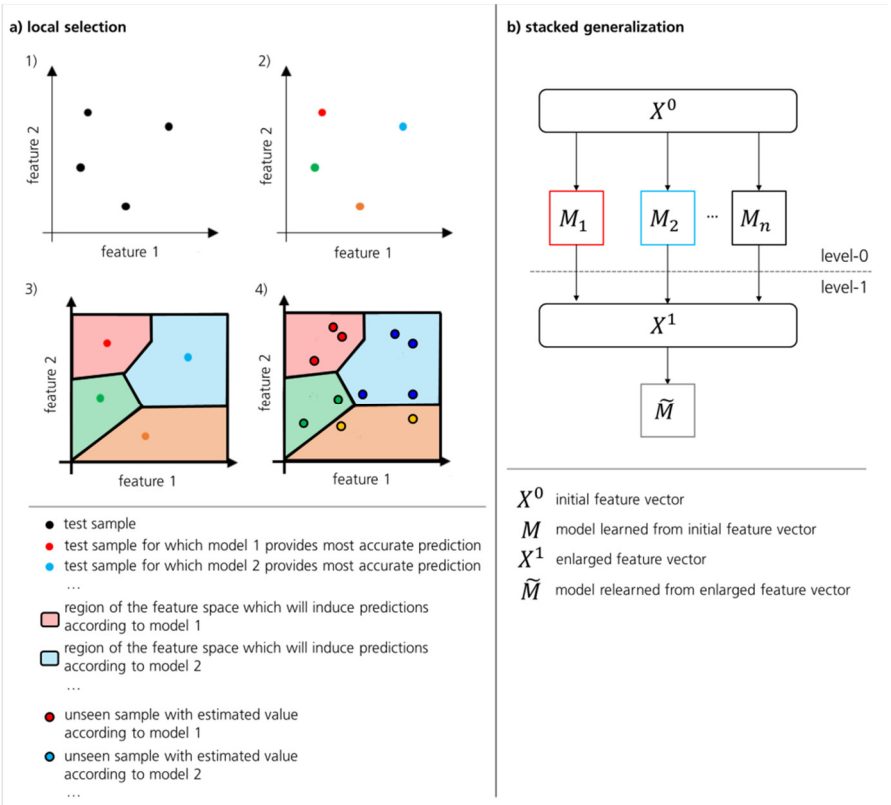

Fig. 2. Scheme of local selection and stacked generalization. (a) Local selection: 1) Exemplification of local selection based on four test samples in a two-dimensional feature space. 2) The regression model which provides the most accurate prediction is identified and assigned to each sample. 3) Partition of the feature space according to the *k*-NN rule. 4) Unseen samples are located in the feature space and predictions are made by the regression model associated to the corresponding region. (b) Stacked generalization: models are learned from an initial training set (level–0). The outputs are stacked to the initial training set to compile a level–1 feature vector, which is used to learn a new model.

2) Stacked generalization

Stacked generalization is a meta-learning approach which deploys the outputs of previously learned models for learning a new model (Wolpert, 1992; Ting and Witten, 1999). As such, the initial model outputs are treated as new features and stacked to the initial feature vector. According to the terminology introduced by Wolpert (1992), the models learned from the initial feature vector correspond to level–0 models and data, respectively (**Fig. 2b**). The enlarged feature vector and the relearned model are called level–1 data and generalizer, respectively. Here, we consider all four regression algorithms separately as level–1 generalizer for relearning the model. Thereby, each level–1 generalizer was learned from a stacked feature vector which includes the outputs of all four level–0 models as established by the four different regression algorithms.

2.6 Domain adaptation

To render our approach robust in an automated manner under covariate shift (i.e., compensate for changes related to acquisition conditions such as illumination and acquisition angle), we include an unsupervised domain adaptation procedure when learning on a source domain (i.e., over a certain city) to predict on a target domain (i.e., over a different city). In particular, we implement a histogram matching procedure regarding the imagery covering source and target domain. This method works according to the principle of adapting data distributions, i.e.,

adapting representations of the original data regardless of the subsequent processing model. The latter point is in particular relevant for our work since we rely on multiple models learned independently. As such, it implements a relative normalization, i.e., it does not provide physical units as an output but similarly distributed digital numbers (Tuia *et al.*, 2016). To that purpose, a nonlinear transformation aligns the shapes of the cumulative histograms of the bands of the image which covers the target domain to the bands of the image which covers the source domain (Gonzalez and Woods, 2002). Subsequently, features are computed from the aligned imagery. Finally, the regression algorithms and the boosting procedure are applied to the source domain and predictions are concatenated with the decision fusion strategies over the target domain.

3. Data and Experimental Setup

3.1 Data

Optical Sentinel-2 data were subject to atmospheric corrections within the Sentinel Application Platform using the Sen2Cor module (ESA, 2018) to provide level 2A products. The imagery for the four cities under investigation was acquired in autumn and winter of the years 2015-2016. The dates were chosen to reduce the influence of photosynthetically active vegetation on the analysis since intra-urban vegetation frequently obscures underlying built-up structures, which then remain undetectable in the corresponding imagery.

The TDM elevation model can be dominantly regarded as a DSM, especially when analyzing built environments as in this study. Only few surfaces such as ice, snow, or vegetation can be penetrated by the X-band SAR signal (Wessel *et al.*, 2018). Comparisons to ICESat data underline the high quality of elevation measurements, which feature less than one meter deviation in absolute vertical accuracy for surfaces other than highly vegetated areas or snow-/ice-covered regions (Rizzoli *et al.*, 2017). Overall, 8 TDM tiles (1° by 1°) with a spatial resolution of 0.4 arcseconds (i.e., ~12 meters) were processed to consistently cover the settlement areas of the four considered cities.

As described previously, two different strategies are considered in this work to establish training data with respect to *built-up density* and *height*.

1) Training set generation from TDM/Sentinel-2 data:

The first strategy builds upon the TDM and Sentinel-2 data to map *built-up density* and *height* with an automatic workflow (Geiß *et al.*, 2017a, 2019). First, it foresees the distinction of "built-up" and "non built-up" areas by relying on the so-called GUF processor (Esch *et al.*, 2012, 2017). This information is subsequently deployed within a tailored filtering procedure for the TDM DSM data to extract elevation information, i.e., compute a normalized DSM (nDSM) for "built-up" areas (Geiß *et al.*, 2015). Finally, the intra-urban land cover of "built-up" areas is mapped under consideration of Sentinel-2 imagery in terms of three thematic classes: "intra-urban vegetation", "elevated built-up", and "residual intra-urban land cover". Consequently, the class "elevated built-up" serves as basis to compute *built-up density* and *height*. In particular, *built-up density* and *height* per grid cell are calculated as follows:

$$built-up\ density_{grid\ cell} = \frac{A_{elevated\ built-up}}{A_{built-up}} \quad (16)$$

where $A_{elevated\ built-up}$ is the area covered by pixels labeled as land cover class "elevated built-up" and $A_{built-up}$ is the area covered by pixels labeled as "built-up" (i.e., GUF);

$$built-up\ height_{grid\ cell} = Q(nDSM_{elevated\ built-up}) \quad (17)$$

where Q is an aggregation function and $nDSM_{elevated\ built-up}$ are the numerical height values contained in the nDSM model for the pixels labeled as land cover class "elevated built-up". For the experiments, we deploy the decile Q_{90} as aggregation function to account for underestimations which are associated with the nDSM data (Geiß *et al.*, 2019).

2) Training set generation from cadastral sources:

As an alternative, we incorporated LoD-1 building geometries and affiliated height measurements, which are based on cadastral information for the four German cities (Wurm *et al.*, 2014). Buildings are represented by extruded footprints in LoD-1 resolution (Luebke *et al.*, 2002). *Built-up density* and *height* per grid cell are computed as in eq. (16) and eq. (17) by substituting the land cover class "elevated built-up" with the LoD-1 building models.

3.2 Experimental Setup and Parameterization

In order to comprehensively describe built-up structures in the imagery, we considered 13 ascending sizes of a square-shaped SE, i.e., $S = \{3,5,7,\dots,21,31,41,51\}$ for the computation of the morphological profiles. Given the spatial resolution of the Sentinel-2 imagery, the morphological operations span over image objects of 30-510m, which is intended to cover various built-up structures. Likewise, the same sizes were deployed for the moving windows, which enable computation of texture measures.

Regarding learning of models, training and test data were strictly spatially separated to avoid biased estimates, which can particularly occur when using spatial features due to the encoding of extrinsic spatial autocorrelation (Geiß *et al.*, 2017c). Experimental results are provided as a function of the size of the spatial processing units. Here, we consider edge length of grid cells according to a linear progression of $a = \{200m, 500m, 800m\}$. Those numerical values allowed reflecting areas of homogeneous urban morphology, i.e., neighborhoods, previously (e.g., Taubenböck *et al.*, 2016). Additionally, for the comparative evaluation of the regression algorithms (**Fig. 3**), we learn models with randomly drawn 100, 150, 200, and 285 samples from the pool of labeled samples. The maximum number of samples considered here is determined by the minimum number of labeled samples available for a settlement area when relying on the maximum size for a. Estimated generalization capabilities are reported as an average of 20 independent trials. For the remaining comparative evaluations (**Fig. 4-7**), we deploy solely the maximum number of samples for model learning.

For the RFR model we tuned the hyperparameters as follows: $n_{tree} = 20,22,\dots,\mathrm{X}$; $m_{try} = 1,2,\dots,\hat{F}$. For the SVR model we deployed Gaussian RBF kernels, which take the form

$K(\mathbf{x}_i\,\mathbf{x}_j) = \exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/2\sigma^2)$. Learning the most appropriate SVR model in conjunction with an RBF kernel requires the definition of the regularization parameter C, the tolerance value ε, and the kernel parameter σ. Model parameters were optimized in the ranges σ = {0.01, 0.02, …, 0.1}, C = {5, 6, …, 15}, and ε = {0.05, 0.06, …, 0.15}. Learning of a GPR model frequently requires exhaustive tuning of hyperparameters related to the covariance or kernel functions. This includes for a RBF kernel the magnitude as well as characteristic expressions of length and noise variance. Consequently, we optimized the model according to σ = {0.01, 0.02, …, 0.1} and ε = {0.001, 0.002, …, 0.1}. Optimization of the NN requires proper regularization of the weights, shape of the nonlinear function, learning rate, as well as model regularization to prevent overfitting. Moreover, the training algorithm and loss function will impact the model. In this manner, the NN was learned based on the RSNNS hyperparameter optimization module (Bergmeier and Benitez, 2012). Model selection was carried out using the RMSE for all algorithms.

For the boosting procedure ϕ was set to 0.05, the number of iterations T was set to 100, and the size of the feature subspace s was set to 10. Regarding the latter it can be noted that s was drawn from the reduced feature vector $\hat{F}$ for a city. Lastly, models are pruned with a σ value of 0.2, which means that only the best 20% of models are considered for the final output.

Regarding the local selection approach, we deployed a k-NN search considering five neighbors. Thereby, the cosine distance measure was used to quantify distances in the feature space. Regarding the stacked generalization method, hyperparameter tuning of level–1 generalizers was carried out the same way as for the level–0 models.

4. Experimental Results and Discussion

Experimental results are computed from test sites which cover the main settlement areas of the four largest German cities: Berlin, Hamburg, Munich, and Cologne. Results are presented as a function of three different learning-validation strategies. The first strategy foresees learning the models with the training set generated from TDM/Sentinel-2 data and provide validation based on the external reference data (i.e., cadastral sources). This corresponds to the most challenging experiment since the validation is based on external reference data with respect to the training set. The second strategy foresees model learning and model validation based on TDM/Sentinel-2 data. This setting aims at answering the question how well the TDM data can be substituted, i.e., how close can Sentinel-2-based predictions resemble the joint TDM/Sentinel-2-based estimations with associated specific data properties and processing principles (Geiß *et al.*, 2019). Lastly, it is also evaluated which levels of accuracy can be achieved when building upon the external reference for model learning and validation.

First, **Fig. 3** is intended to decompose the consecutive gain in accuracy when building a multistrategy ensemble as proposed. It provides boxplots which document the obtained mean absolute errors (MAE) when using the single regression algorithms, boosted regression algorithms, and models which additionally internalize the decision fusion strategies with respect to *built-up density* (**Fig. 3a**) and *height* (**Fig. 3b**) for different sizes of the grid cells and learning-validation strategies, respectively.

Regarding the applied MCFS feature selection algorithm, it can be noted that the composition of feature subsets varied a lot. Depending on the target variable, validation strategy, size of the spatial processing units, and test area, the feature vector used for model learning is reduced to a dimensionality between 20 and 51 based on the initial feature vector (with 663 dimensions). Thereby, only the normalized difference between the red and green band ($normalized\ red\ green\ index_{mean}$) from the spectral feature category is always selected for the feature subset when estimating *built-up density* in our experiments.

In general, it can be observed that the MAE decreases for all four test cities with an increasing size of the grid cells throughout the different learning-validation strategies, which can be expectedly related to the central limit theorem. Regarding the *built-up density* estimations, unsurprisingly, the largest deviations can be observed when learning the models with the training set generated from TDM/Sentinel-2 data and provide validation based on the cadastral sources. This can be particularly related to the fact that the cadastral sources contain high *built-up density* not only in the central parts of the city but also in fringe areas, which is hardly the case for the TDM/Sentinel-2-based estimates (Geiß *et al.*, 2019). However, the remaining results, including accuracies obtained from the *built-up height* estimations, show comparable accuracy levels independent of the learning-validation strategy, which reflects generally viable model estimates. From a comparative model perspective, it can be noted that the single regression algorithms can be consecutively enhanced when internalizing the boosting strategy and finally also the decision fusion strategies. Corresponding median values of predictions decrease unambiguously. Regarding all configurations shown in **Fig. 3**, median values of predictions decrease with respect to the single regression algorithms on average by 7.6 and 12.9 percentage points (p.p.) as well as 7.9 and 14.1 p.p. regarding the boosted regressors and decision fusion strategies for *built-up density* and *height*, respectively, which clearly indicates the beneficial performance properties of the MSER method.

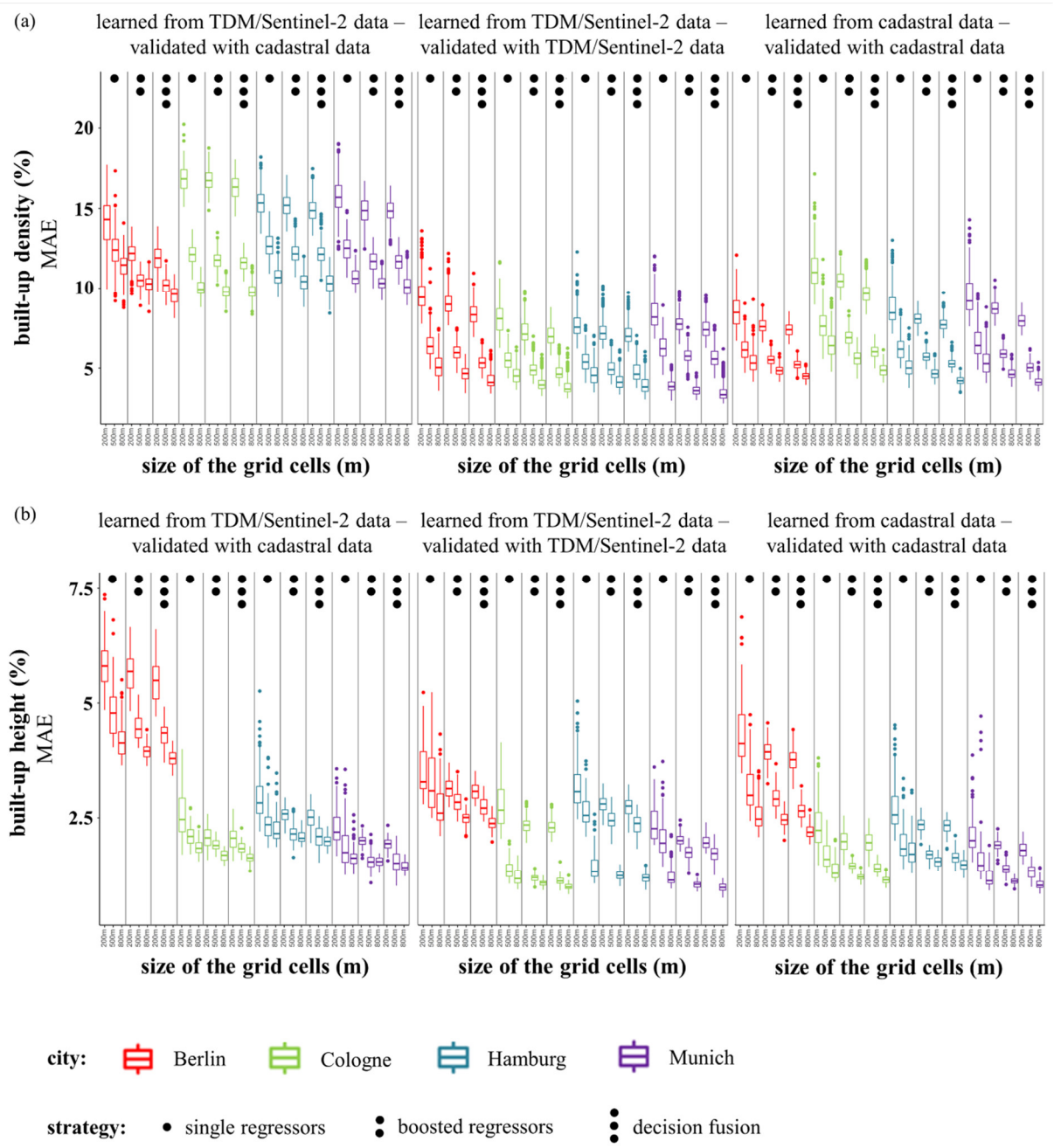

Fig. 3. Boxplots represent accuracies in terms of MAE obtained with the single regression algorithms, boosted regression algorithms, and models which additionally internalize the decision fusion strategies with respect to (a) *built-up density* and (b) *height* for different sizes of the grid cells (i.e., 200m, 500m, 800m) and learning-validation strategies, respectively.

To evaluate the accuracy levels of solely the best predictions, **Fig. 4a** and **Fig. 4b** provide mean error (ME) and MAE with respect to *built-up density* and *height* based on the best models, respectively. Unsurprisingly and in concordance with previous findings, ME values for *built-up density* show highest levels when learning the models with the training set generated from TDM/Sentinel-2 data and provide validation based on external data (i.e., cadastral sources). The negative ME values indicate here an underestimation of *built-up density*. In contrast, the ME shows almost zero deviations with respect to the learning strategies which do not foresee a validation based on external data. In combination with non-zero MAE values, this indicates a balance of over- and underestimations and, simultaneously,

a valid fit of the models. As such, MAE values show values between 10% and 16% for models learned with the training set generated from TDM/Sentinel-2 data and validated based on the cadastral sources. Even more favorable accuracy levels can be achieved when relying on non-external data for model learning and validation: corresponding MAE values range from 3% to 9%. The *built-up height* estimations show a very similar accuracy pattern in terms of ME and MAE regarding the different learning-validation strategies. Thereby, predictions regarding the incorporation of external validation data feature a ME from -4.5m to 0.4m and a MAE from 5.4m to 1.4m. In addition, predictive accuracies computed without the incorporation of external validation data show values between -0.5m and 0.01m as well as between 3.8m and 1m with respect to ME and MAE, respectively. This underlines the viability to learn predictive models also for *built-up height* estimations based on Sentinel-2 features. However, generally it should be noted that deviation levels are strongly influenced by the morphologic structure of a city: larger numerical values in terms of *built-up density* and *height* feature larger deviations (Geiß *et al.*, 2019). Besides, the results unambiguously underline the beneficial properties of the MSER approach, since all best predictions were obtained with a boosted regressor in conjunction with a decision fusion strategy. In this manner, 23 and 8 out of the 36 best predictions for *built-up density* were obtained with the stacked generalization step and NN as level–1 generalizer, and the local selection strategy, respectively. Also, for *built-up height* estimations, stacked generalization models with NN as level–1 generalizer underlined their beneficial performance properties by providing 17 times the best prediction. Equally often, stacked generalization models with SVR as level–1 generalizer could provide the best prediction.

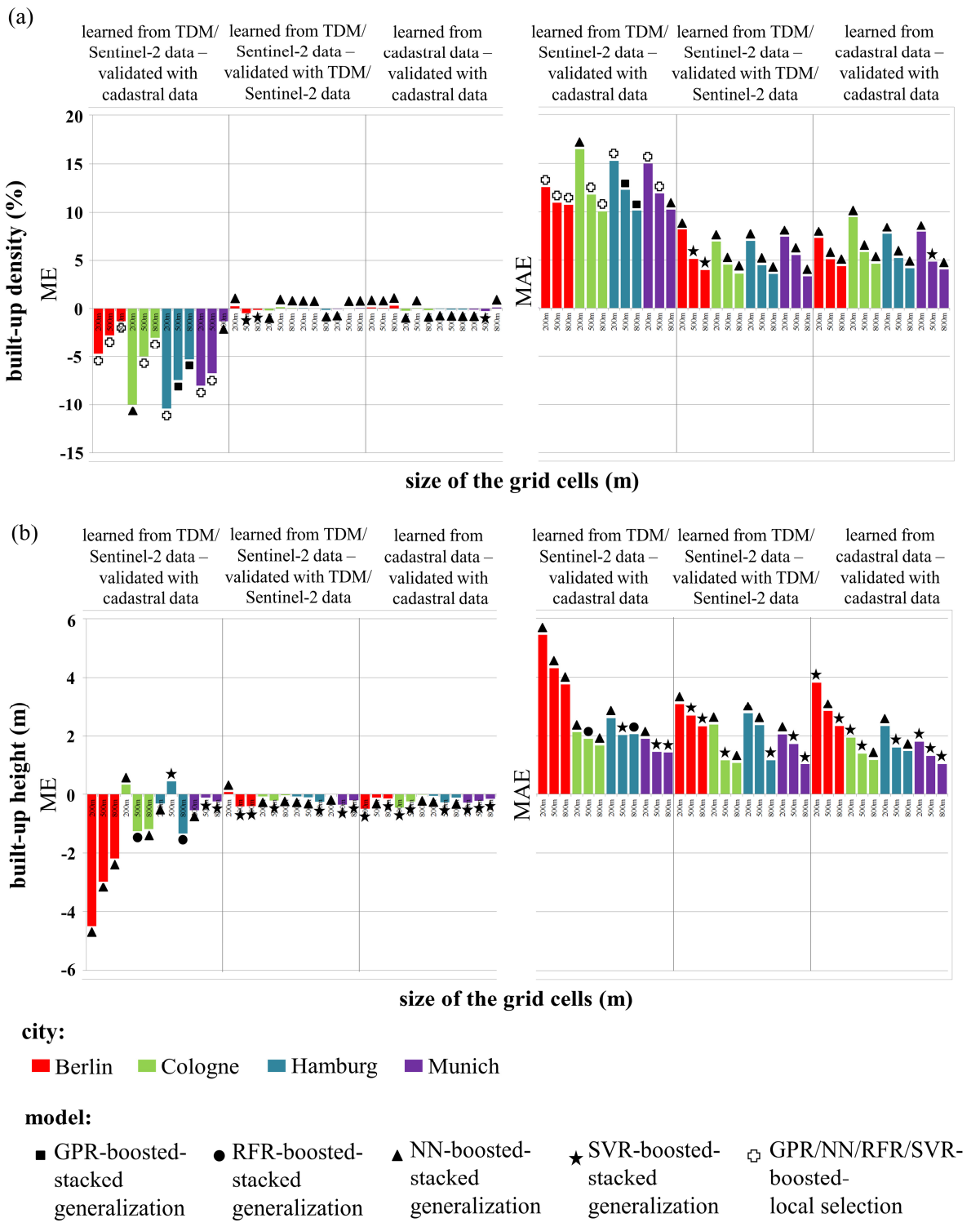

Fig. 4. Accuracies of best model predictions in terms of ME and MAE for the four test sites as a function of the learning-validation strategies and different sizes of the grid cells (i.e., 200m, 500m, 800m) for (a) *built-up density* and (b) *height*, respectively.

To illustrate further, **Fig. 5a** and **Fig. 5b** provide a relative spatial visualization of the best model results for comparing the different learning-validation strategies with respect to *built-up density* and *height*, respectively, for the city of Berlin. Regarding *built-up density* mapping, the training data based on both TDM/Sentinel-2 data and cadastral sources show an ideal decrease from the core to the fringe of the city. Thereby, *built-up density* derived from the cadastral sources show also highest values in the urban core. However, the values are more spatially fragmented. In addition, also high values can be found in vast parts of the fringe areas. Nevertheless, the spatial pattern of both training data sets can be very well reflected by the model estimates. Analogously, the training data for *built-up height* show the urban core with *high built-up height* values and *low built-up height* values are characteristic for the peripheral areas of the city for both data sources. Thereby, cadastral data feature an ever more compact pattern. Model estimates for both data sets are able to trace this spatial distribution of *built-up height* very clearly. As such, this analysis further underlines the property of Sentinel-2 imagery to carry the morphologic characteristics of built environments.

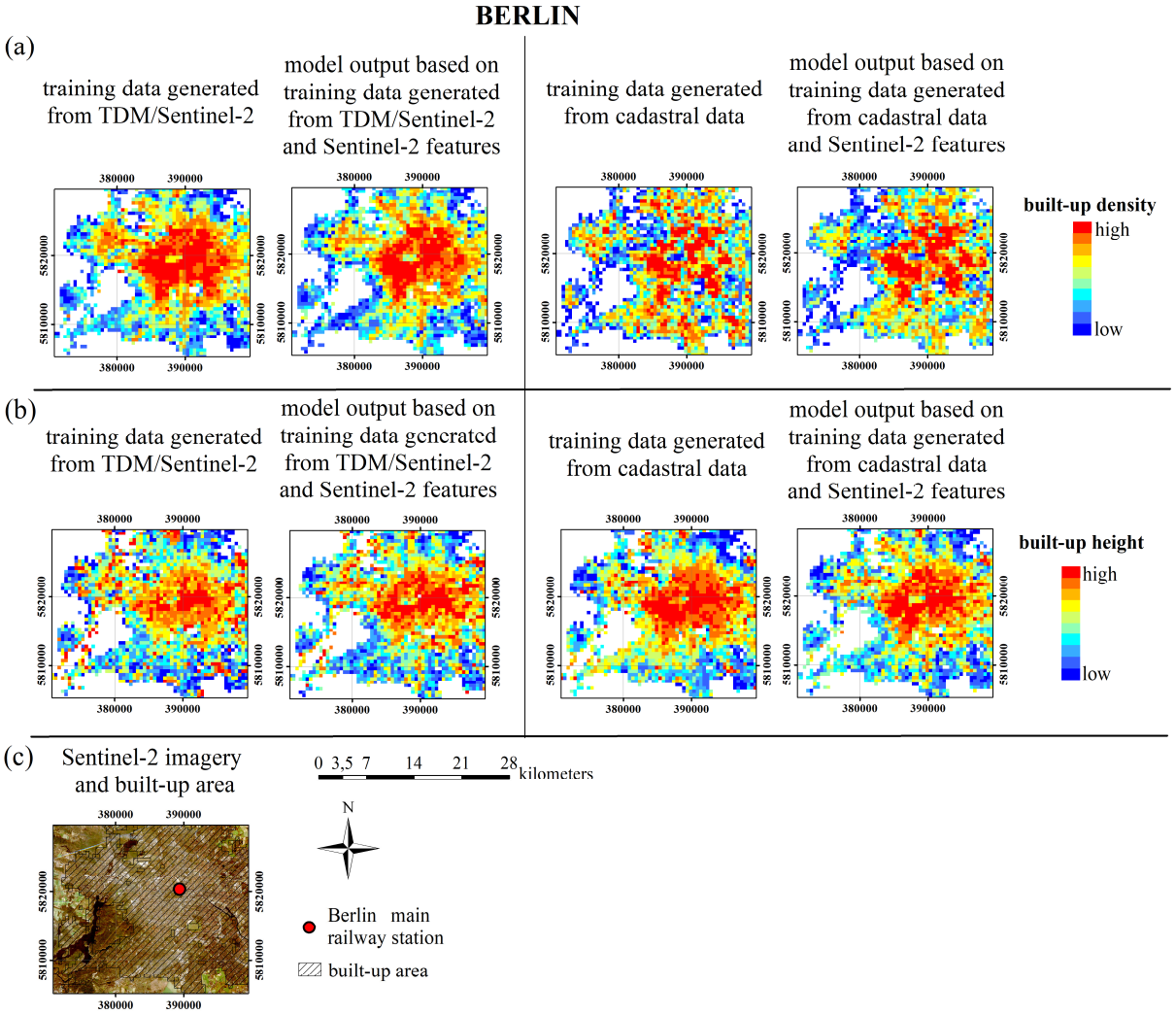

Fig. 5. Visualized (a) *built-up density* and (b) *height* according to a = 500m and the different learning-validation strategies for the example of Berlin, Germany. The color coding corresponds to deciles for the individual data sets and, thus, allows for a relative spatial comparison. (c) Sentinel-2 imagery with superimposed built-up area and main railway station of Berlin for orientation.

Lastly, obtained accuracies in terms of MAE from the decision fusion strategies in the domain adaptation experiments are documented in **Fig. 6a** and **Fig. 6b** with respect to *built-up density* and *height*, respectively, as a function of size of the grid cells, learning-validation strategy, and source domain. Analogous to previous results, accuracy levels of *built-up density* estimations mainly increase with an increasing size of the grid cells throughout the different learning-validation strategies. Thereby, the results obtained with the training set generated from TDM/Sentinel-2 data and the validation set based on the cadastral sources feature a comparable accuracy level as the results obtained within a source domain (**Fig. 6a**). When relying on non-external data for model learning and validation, accuracies can feature lower levels compared to accuracies achieved within a source domain. Generally, accuracy levels differ significantly depending on the combination of source and target domain. This holds true also for *built-up height* estimations, where most favorable combinations of source and target domain allow for similar accuracy levels as from estimations within a source domain (**Fig. 6b**). Here, the city of Berlin shows the lowest correspondence with respect to the other cities (i.e., when Berlin is the source domain, accuracies are comparably low for the other cities and the accuracies for Berlin are comparably low regardless of which other city was selected as source domain), which indicates that the morphology of Berlin's *built-up height* is significantly different with respect to the other cities and further underlines the need for a proper choice of the combination of source and target domain. Nevertheless, numerous predictions with high accuracies underline the viability to transfer a model and, thus, enable a substitution of the training data in the target domains. Thereby, model accuracies increase on average by 9.9 p.p. when deploying the cadastral sources for learning and validation compared to the accuracies achieved when learning with TDM/Sentinel-2 data and validating with cadastral sources. This difference can be interpreted as an error cost when considering integrating data with reference quality (here from cadastral sources) compared to the more ubiquitously available TDM/Sentinel-2-based estimates.

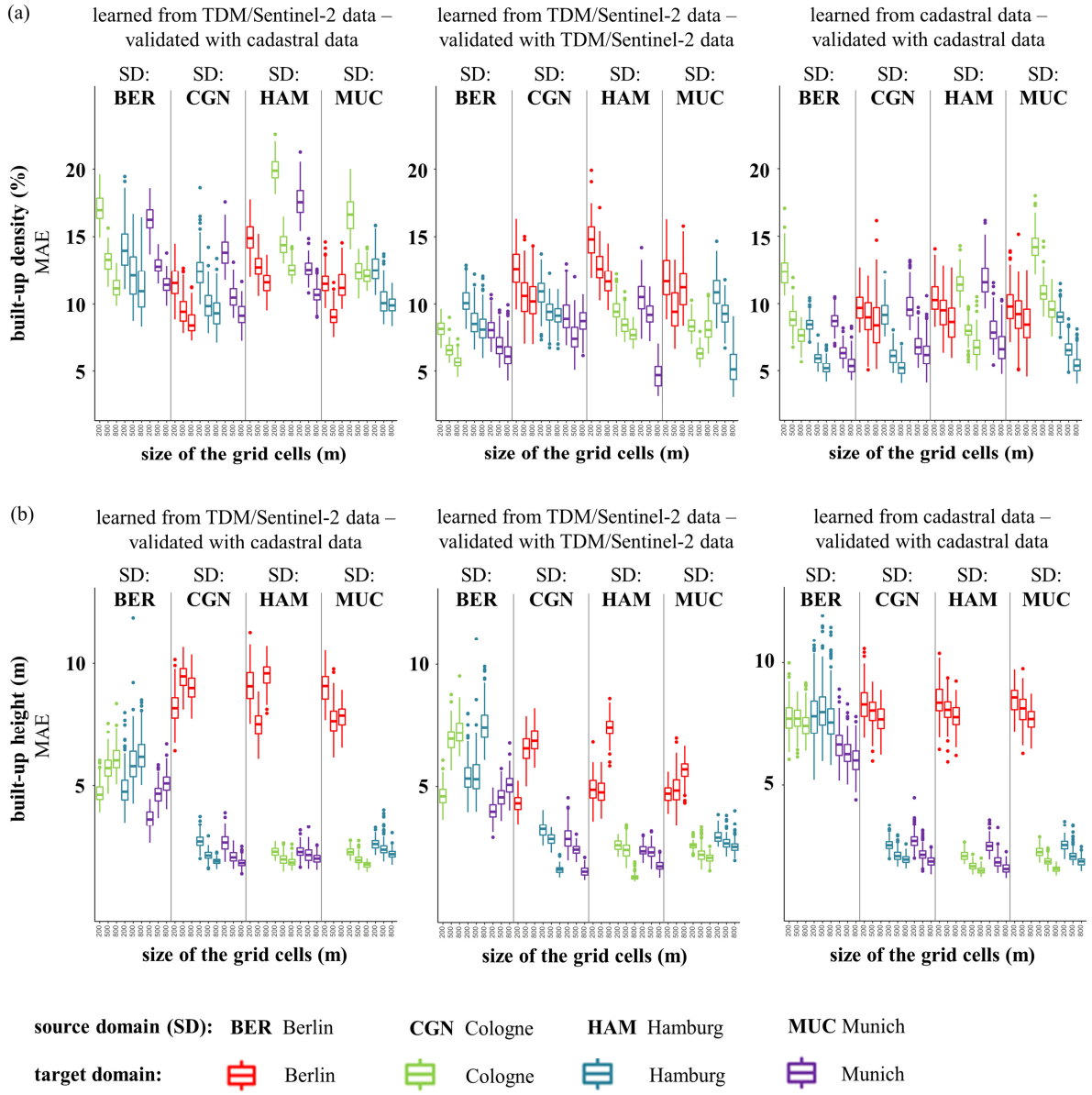

Fig. 6. Boxplots represent accuracies in terms of MAE from the decision fusion strategies in the domain adaptation experiments with respect to (a) *built-up density* and (b) *height* for different sizes of the grid cells (i.e., 200m, 500m, 800m), learning-validation strategies, and source domain, respectively.

5. Concluding remarks and future perspectives

We have proposed a novel ensemble regression approach, i.e., MSER, to estimate *built-up density* and *height* with features derived from Sentinel-2 data. Experimental results uncovered the beneficial performance properties of the MSER method, since it consistently provided most accurate model predictions in a comparative setup. The main findings are:

1) From the exhaustive initial feature vector (with 663 dimensions), which contains features from the spectral, mathematical morphology, and texture domain, only a substantially reduced subset (with a dimensionality between 20 and 51) was compiled by the MCFS feature selection algorithm and deployed for regression. Thereby, only the normalized difference between the red and green band was always selected in this study for estimating *built-up density*.

2) From a comparative model perspective, the individual regression algorithms (i.e., RFR, SVR, GPR, NN) can be enhanced step by step when also following the proposed boosting strategy and finally also implementing the decision fusion strategies. MAE values of regression estimates improve with respect to the single regression algorithms on average by 7.6 and 12.9 p.p. for *built-up density* as well as 7.9 and 14.1 p.p. for *built-up height* regarding the boosted regressors and decision fusion strategies, respectively.

3) For the estimation of *built-up density*, it turned out that majority of best accuracies could be obtained with stacked generalization models, whereby NN served as level–1 generalizer. For the estimation of *built-up height*, stacked generalization models with NN and SVR as level–1 generalizer, respectively, allowed equally often to obtain the best accuracy.

4) Finally, the viability to transfer a model was shown in a domain adaptation context. However, experimental results also uncovered that a proper choice of the combination of source and target domain is required to maintain viable accuracy levels of the regression estimates.

Overall, the accuracy levels of the regression estimates are very promising. The mean absolute errors of the models varied in this study between 3–16% and 1–5.4m with respect to *built-up density* and *height* estimates, respectively, depending on the learning-validation strategy, size of the spatial processing units, and test area. Thereby, from a relative spatial distribution perspective, model estimates are able to trace the relative spatial configuration of *built-up density* and *height* over the test areas very clearly. Nevertheless, in the future, we aim for enhancing the applicability of the approach by designing an unsupervised method which automatically selects the best combination of source and target domains to minimize problems related to sample selection bias. Moreover, future work can also exploit multioutput regression models for a likely beneficial joint estimation of the two target variables considered here.

Acknowledgements

The work of Christian Geiß was supported by the Helmholtz Association under the grant “pre_DICT” (PD-305). This research was also funded in part by the German Federal Ministry of Education and Research (BMBF) under grant no. 03G0876 (project RIESGOS). The authors also would like to thank the German Research Foundation (DFG) for financing the research project “Where are the jobs? Stadtregionale Zentrenstrukturen im internationalen Vergleich” with the grant number: TA 800/6-1 & SI 932/12-1. We also want to thank the anonymous reviewers for very helpful comments on the initial version of the paper.

References

H. Aghighi *et al.*, “Machine Learning Regression Techniques for the Silage Maize Yield Prediction Using Time-Series Images of Landsat 8 OLI,” *IEEE J. Sel. Topics Appl. Earth Observations Remote Sens.*, vol. 11, no. 12., 2018.

P. Aravena Pelizari *et al.*, “Multi-sensor feature fusion for very high spatial resolution built-up area extraction in temporary settlements,” *Remote Sens. Environ.*, vol. 209, pp. 793–807.

C. Berger *et al.*, "Multi-modal and multi-temporal data fusion: Outcome of the 2012 GRSS data fusion contest," *IEEE J. Sel. Topics Appl. Earth Observations Remote Sens.*, vol. 6, no. 3, pp. 1324– 1340, Jun. 2013.

C. Bergmeir and J. M. Benitez, "Neural Networks in R Using the Stuttgart Neural Network Simulator: RSNNS," Journal of Statistical Software, vol. 46, no. 7, pp. 1–26, 2012

L. Breiman, "Bagging predictors," *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, 1996.

L. Breiman, "Statistical modeling: The two cultures," *Statist. Sci.*, vol. 16, no. 3, pp. 199–231, 2001.

L. Breimann, "Random Forest," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

L. Bruzzone and F. Melgani, "Robust Multiple Estimator Systems for the Analysis of Biophysical Parameters From Remotely Sensed Data," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 43, no. 1, pp. 159–174, 2005.

X. Chen *et al.*, "Supervised Multiview Feature Selection Exploring Homogeneity and Heterogeneity With ℓ1,2-Norm and Automatic View Generation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 4, pp. 2074–2088, 2017.

G. Camps-Valls *et al*., "A survey on Gaussian Processes for Earth Observation Data Analysis – A comprehensive investigation," *IEEE Geoscience and Remote Sensing Magazine*, 2016.

Copernicus - Copernicus Open Access Hub; URL: https://scihub.copernicus.eu/ Last accessed: 27 Aug 2018.

C. Cortes and V. Vapnik, "Support vector networks," *Mach. Learn.*, vol. 20, pp. 273–297, 1995.

M. Drusch *et al.*, "Sentinel-2: ESA's Optical High-Resolution Mission for GMES Operational Services," *Remote Sens. Environ.*, vol. 120, pp. 25–36, May 2012.

P. Du *et al.*, "Multiple classifier system for remote sensing image classification: a review," *Sensors*, vol. 12, pp. 4764–4792, 2012.

W. Duch, "Filter methods," In: Guyon, I., Nikravesh, M., Gunn, S., Zadeh, L.A. (eds). Feature Extraction, pp. 89–117, 2006.

European Space Agency (ESA) – Sen2Cor; URL: http://step.esa.int/main/third-party-plugins-2/sen2cor/ Last accessed: 19 May 2018.

T. Esch *et al.*, "Large-area assessment of impervious surface based on integrated analysis of single-date Landsat-7 images and geospatial vector data," *Remote Sensing of Environment*, vol. 113, no. 8, pp. 1678–1690, 2009.

T. Esch et al., "TanDEM-X mission: New perspectives for the inventory and monitoring of global settlement patterns," *Journal of Applied Remote Sensing*, vol 6, no. 1, 21 pages.

T. Esch *et al.*, "Breaking new ground in mapping settlements from space – The Global Urban Footprint," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 134, pp. 30–42, 2017.

H. Feilhauer *et al.*, "Multi-method ensemble selection of spectral bands related to leaf biochemistry," *Remote Sens. Environ.*, vol. 164, pp. 57–65, 2015.

Y. Freund and R. E. Shapire, "Experiments with a new boosting algorithm," in Proc. 13th Int. Conf. Mach. Learn., Bari, Italy, 1996, pp. 1–9.

Y. Freund and R. E. Schapire, „A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of Computer and System Sciences*, vol. 55, no. 1, pp. 119-139, 1997.

N. García-Pedrajas and D. Ortiz-Boyer, "Boosting random subspace method," *Neural Networks*, vol. 21, no. 9, pp. 1344–1362, 2008.

C. Geiß and H. Taubenböck, "Remote sensing contributing to assess earthquake risk: From a literature review towards a roadmap," *Natural Hazards*, vol. 68, no. 1, pp. 7–48, Aug. 2013.

C. Geiß *et al*., "Remote sensing-based characterization of settlement structures for assessing local potential of district heat," *Remote Sens.*, vol. 3, no. 7, pp. 1447–1471, 2011.

C. Geiß *et al.*, "Estimation of seismic buildings structural types using multi-sensor remote sensing and machine learning techniques," *ISPRS J. Photogramm. Remote Sens.*, vol. 104, pp. 175–188, 2015.

C. Geiß *et al.*, "Estimation of Seismic Vulnerability Levels of Urban Structures With Multisensor Remote Sensing," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 9, no 5, pp. 1913-1936, May 2016.

C. Geiß *et al.*, "Object-based Morphological Profiles for Classification of Remote Sensing Imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 10, pp. 5952-5963, 2016.

C. Geiß *et al.* (2017a) Towards large-area morphologic characterization of urban environments using the TanDEM-X mission and Sentinel-2, *2017 Joint Urban Remote Sensing Event (JURSE)*, Dubai, United Arab Emirates, pp. 1-4. doi: 10.1109/JURSE.2017.7924543.

C. Geiß *et al.*, "Multitask Active Learning for Characterization of Built Environments With Multisensor Earth Observation Data," *IEEE J. Sel. Topics Appl. Earth Obs. Remote Sens.*, vol. 10, no. 12, pp. 5583-5597, 2017b.

C. Geiß *et al.*, "On the Effect of Spatially Non-disjoint Training and Test Samples on Estimated Model Generalization Capabilities in Supervised Classification with Spatial Features," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 11, pp. 2008-2012, 2017c.

C. Geiß, *et al.* "Large-Area Characterization of Urban Morphology – Mapping of Built-Up Height and Density Using TanDEM-X and Sentinel-2 Data," *IEEE J. Sel. Topics Appl. Earth Obs. Remote Sens*, vol. 12, no. 8, pp. 2912-2927, 2019.

C. Geiß *et al.*, "Normalization of TanDEM-X DSM data in urban environments with morphological filters," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 8, pp. 4348–4362, Aug. 2015.

P. Ghamisi and N. Yokoya, "IMG2DSM: Height Simulation From Single Imagery Using Conditional Generative Adversarial Net," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 5, pp. 794–798.

G. Giacinto and F. Roli, "Dynamic classifier selection based on multiple classifier behavior," *Pattern Recognit*., vol. 34, pp. 1879–1881, 2001.

R. C. Gonzalez and R. E. Woods, Digital Image Processing, 2nd ed. Upper-Saddle River, NJ, USA: Prentice-Hall, 2002.

N. Gorelick *et al.* "Google Earth Engine: Planetary-scale geospatial analysis for everyone," *Remote Sensing of Environment*, vol. 202, pp. 18–27, 2017.

I. Guyon, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1157–1182, 2003.

M. A. Hall, "Correlation-based Feature Selection for Discrete and Numeric Class Machine Learning," In: Proceedings of the Seventeenth International Conference on Machine Learning (ICML '00), Pat Langley (Ed.). Morgan Kaufmann Publishers Inc., San Francisco, pp. 359–366, 2000.

R. M. Haralick, "Statistical and structural approaches to texture," *Proc. IEEE*, vol. 67, no. 5, pp. 786–804, 1979.

U. Heiden *et al.*, "Urban structure type characterization using hyperspectral remote sensing and height information," *Landscape and Urban Planning*, vol. 105, no. 4, pp. 361–375, 2012.

J. Heinzel and T. Kemper, "Automatic metric characterization of urban structure using building decomposition from very high resolution imagery," *International Journal of Applied Earth Observation and Geoinformation*, vol. 35, pp. 151–160, 2015.

T. K. Ho, "The random subspace method for constructing decision forests," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 8, pp. 832–844, 1998.

X. Huang and L. Zhang, "Morphological Building/Shadow Index for Building Extraction from High-resolution Imagery Over Urban Areas," *IEEE J. Sel. Topics Appl. Earth Obs. Remote Sens*., vol. 5, no. 1, pp. 161-172, 2012.

X. Huang *et al.*, "A multi-index learning approach for classification of high-resolution remotely sensed images over urban areas," *ISPRS J. Photogramm. Remote Sens.*, vol. 90, pp. 36–48, 2014.

X. Huang *et al.*, "Multi-level monitoring of subtle urban changes for the megacities of China using high-resolution multi-view satellite imagery," *Remote Sensing of Environment*, vol. 196, pp. 56–75, 2017.

X. Huang et al., "Angular difference feature extraction for urban scene classification using ZY-3 multi-angle high-resolution satellite imagery," *ISPRS J. Photogramm. Remote Sens.*, vol. 135, pp. 127–141, 2018.

G. Hughes, "On the mean accuracy of statistical pattern recognizers," *IEEE Transactions on Information Theory*, vol. 14, no. 1, pp. 55–63, 1968.

R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artif. Intell.*, vol. 97, no. 1, pp. 273–324, 1997.

G. Krieger *et al.*, "TanDEM-X: A satellite formation for high-resolution SAR interferometry," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 11, pp. 3317–3341, Nov. 2007.

N. Kummer and H. Najjaran, H., "Adaboost.MRT: Boosting regression for multivariate estimation," *Artif. Intell. Research*, vol. 3, no. 4, pp. 64-76, 2014.

T. N. Lal *et al.*, "Embedded methods," In: Guyon, I., Nikravesh, M., Gunn, S., Zadeh, L.A. (eds). Feature Extraction, pp. 137–165.

A. Lefebvre *et al.*, "Monitoring Urban Areas with Sentinel-2A Data: Application to the Update of the Copernicus High Resolution Layer Imperviousness Degree," *Remote Sensing*, vol. 8, 606, doi:10.3390/rs8070606 , 2016.

P. Leinenkugel *et al.*, "Settlement detection and impervious surface estimation in the Mekong Delta using optical and SAR remote sensing data," *Remote Sensing of Environment*, vol. 115, no. 12, pp. 3007–3019, 2009.

A. Liaw and M. Wiener, "Classification and Regression by randomForest," *R News: The Newsletter of the R Project*, vol. 2, no. 3, pp. 18–22, 2002.

C. Liu et al., "Automatic extraction of built-up area from ZY3 multi-view satellite imagery: Analysis of 45 global cities," *Remote Sensing of Environment*, vol. 226, pp. 51–73, 2019.

D. Luebke, B. Watson, J.D. Cohen, M. Reddy, and A. Varshney, Level of Detail for 3D Graphics. Elsevier Science Inc., 2002.

J. Mendes-Moreira *et al.*, "Ensemble Approaches for Regression: A Survey," *ACM Computing Surveys (Csur)*, vol. 45, no. 1, pp. 1–40, 2012.

L. Mou and X.X. Zhu, "IM2HEIGHT: Height estimation from single monocular imagery via fully residual convolutional-deconvolutional network," arXiv:1802.10249.

F. Pacifici *et al.*, "A neural network approach using multi-scale textural metrics from very high resolution panchromatic imagery for urban land-use classification," *Remote Sens. Environ.*, vol. 113, no. 6, pp. 1276–1292, 2009.

M. Pesaresi and J. A. Benediktsson, "A new approach for the morphological segmentation of high-resolution satellite imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 39, no. 2, pp. 309–320, Feb. 2001.

M. Pesaresi and A. Gerhardinger, "Improved textural built-up presence index for automatic recognition of human settlements in arid regions with scattered vegetation," *IEEE J. Sel. Topics Appl. Earth Obs. Remote Sens.*, vol. 4, no. 1, pp. 16–26, Mar. 2011.

M. Pesaresi, "A robust built-up area presence index by anisotropic rotation-invariant textural measure," *IEEE J. Sel. Top. Appl. Earth Observations Remote Sens.*, vol. 1, no. 3, pp. 180–192, 2008.

M. Pittore *et al.*, "Perspectives on global dynamic exposure modelling for geo-risk assessment," *Natural Hazards*, vol. 86, pp. 7–30, 2017.

R. Policar, "Ensemble based systems in decision making," *IEEE Circuits Syst.Mag.*, vol. 6, pp. 21–45, 2006.

C. E. Rasmussen and C. K. I. Williams, Gaussian Processes for Machine Learning. New York, NY, USA: MIT Press, 2006.

C. Ratti *et al.*, "Energy consumption and urban texture," Energy and Buildings, vol. 37, no. 7, pp. 762-776, 2005.

J. W. Rouse, Jr., R. H. Haas, J. A. Schell, and D. W. Deering, "Monitoring vegetation systems in the Great Plains with ERTS," in Proc. 3rd Earth Resour. Technol. Satell. Symp., Washington, DC, USA, 1973, pp. 309–317.

S. J. Russell and P. Norvig, "Artificial Intelligence: A Modern Approach," 3rd Edition, Pearson, Upper Saddle River, 1132 p., 2010.

P. Soille, Morphological Image Analysis: Principles and Applications. Springer, 2004.

D. P. Solomatine and D. L. Shrestha, "AdaBoost.RT: a boosting algorithm for regression problems," in Proceedings of the IEEE International Joint Conference on Neural Networks, vol. 2, pp. 1163–1168, IEEE, 2004.

A. Stumpf and N. Kerle, "Object-oriented mapping of landslides using Random Forests," *Remote Sens. Environ.*, vol. 115, pp. 2564–2577, 2011.

H. Taubenböck *et al.*, "Monitoring urbanization in mega cities from space," *Remote Sens. Environ.*, vol. 117, pp. 162–176, Feb. 2012.

H. Taubenböck *et al.* ,"The Physical Density of the City – Deconstruction o the Delusive Density Measure with Evidence from Two European Megacities," *ISPRS Int. J. Geo-Inf.*, vol. 5, no. 11, 2016.

K. M. Ting and I. Witten, "Issues in Stacked Generalization," *Journal of Artificial Intelligence Research*, vol. 10, pp. 271–289, 1999.

D. Tuia *et al.*, "Domain Adaptation for the Classification of Remote Sensing Data – an overview of recent advances," *IEEE Geoscience and Remote Sensing Magazine*, 2016.

P. Rizzoli *et al.*, "Generation and performance assessment of the global TanDEM-X digital elevation model," *ISPRS J. Photogramm. Remote Sens.*, vol. 132, pp. 119–139, 2017.

A. Okujeni *et al.*, "Ensemble learning from synthetically mixed training data for quantifying urban land cover with support vector regression," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* vol. 10, pp. 1640–1650, 2017.

G. Wang and J. Ma, "Study of corporate credit risk prediction based on integrating boosting and random subspace," Expert Systems with Applications, vol. 38, no. 11, pp. 13871–13878, 2011.

G. I. Webb and Z. Zheng, "Multistrategy Ensemble Learning: Reducing Error by Combining Ensemble Learning Techniques," *IEEE Transactions on Knowledge and Data Engineering*, vol. 16, no. 8, pp. 980–991, 2004.

B. Wessel et al., "Accuracy assessment of the global TanDEM-X Digital Elevation Model with GPS data," *ISPRS J. Photogramm. Remote Sens.*, vol. 139, pp. 171–182, 2018.

H. Wolpert, "Stacked generalization," *Neural Networks*, vol. 5, pp. 241–259, 1992.

H. Wolpert, "The lack of a priori distinctions between learning algorithms," *Neural Computation*, vol. 9, pp. 1341–1390, 1996.

M. Wurm, P. d'Angelo, P. Reinartz, and H. Taubenböck, "Investigating the Applicability of Cartosat-1 DEMs and Topographic Maps to Localize Large-Area Urban Mass Concentrations," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no 10, pp. 4138-4152, Oct. 2014.

J. Verrelst *et al*., "Machine learning regression algorithms for biophysical parameter retrieval: Opportunities for Sentinel-2 and-3," *Remote Sensing of Environment*, vol. 118, pp. 127–139, 2009.

R. Xu *et al.*, "Extraction of High-Precision Urban Impervious Surfaces from Sentinel-2 Multispectral Imagery via Modified Linear Spectral Mixture Analysis," *Sensors*, vol. 18, 2873, doi:10.3390/s18092873, 2018.

B. Yu *et al.*, "Automated derivation of urban building density information using airborne LiDAR data and object based method," *Landscape Urban Plan.*, vol. 98, nos. 3/4, pp. 210–219, Dec. 2010

T. Zhang *et al.*, "Urban Building Density Estimation from High-Resolution Imagery Using Multiple Features and Support Vector Regression," *IEEE J. Sel. Topics Appl. Earth Observations Remote Sens.*, vol. 10, no. 7, pp. 3265–3280, Jul. 2017.

J. Zhao *et al.*, "Multi-view learning overview: recent progress and new challenges," *Inf. Fusion*, vol. 38, pp. 43–54, 2017.

M. Zink *et al.*, "TanDEM-X: the new global DEM takes shape," *IEEE Geosci. Remote Sens. Mag.*, vol. 2, no. 2, pp. 8–23, 2014.